

INTRINSIC DISORDER AND PROTEIN EVOLUTION:
AMINO ACID COMPOSITION OF PROTEINS IN LAST
UNIVERSAL ANCESTOR.

Sai Harish Babu Karne

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Bioinformatics
Indiana University

December 2008

Accepted by the Faculty of Indiana University,
In partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

Master's Thesis Committee

Dr. Pedro Romero, PhD, Chair.

Dr. Keith A Dunker, PhD.

Dr. Vladimir N. Uversky, PhD.

© 2008

Sai Harish Babu Karne

ALL RIGHTS RESERVED

Dedicated to My Parents.

TABLE OF CONTENTS

	Page
LIST OF TABLES	VII
LIST OF FIGURES	VIII
ACKNOWLEDGEMENTS	IX
ABSTRACT	X
CHAPTER ONE: INTRODUCTION	1
1.1 INTRINSICALLY DISORDERED PROTEINS	1
1.2 EVOLUTION	3
1.3 LAST UNIVERSAL ANCESTOR	5
1.4 ESTIMATION OF ANCESTOR PROTEINS	7
CHAPTER TWO: LITERATURE REVIEW	9
2.1 DISORDERED PROTEINS	9
2.2 PROTEIN EVOLUTION	14
2.3 HYPOTHESIS(ES)	17
CHAPTER THREE: METHODS AND MATERIALS	20
3.1 ESTIMATION OF AMINO ACID COMPOSITION OF THE LUA	20
3.2 PHYLOGENETIC TREE	23
3.3 CLUSTALW	24

3.4 PROTEINS DATA OF INTEREST	25
3.5 ENZYMES & NON-ENZYMES.....	28
3.6 APPLICATION OF THE SOFTWARE ON THE DATA SET	28
3.7 FREQUENCY OF ‘C’ ‘F’ ‘W’ ‘Y’	30
3.8 LENGTH BIAS	33
3.9 BOOTSTRAPPING.....	33
 CHAPTER FOUR: RESULTS	 36
4.1 LUA COMPARISONS.....	36
4.2 RESULTS OF THE ANALYSIS ON PROTEIN SEQUENCES.....	43
4.3 SUMMARY OF FINDINGS	54
 CHAPTER FIVE: DISCUSSION.....	 55
5.1 ELUCIDATION OF OUTCOMES	55
5.2 SIGNIFICANCE OF RESULTS	57
5.3 REVIEW OF DISCUSSIONS.....	58
 CHAPTER SIX: CONCLUSIONS.....	 60
6.1 LIMITATIONS IN THE STUDY	60
6.2 FUTURE STUDY.....	60
6.3 FINAL SUMMARY	61
References.....	62
Appendices.....	72
VITA	73

LIST OF TABLES

4.1 LUA OF 65 COGS AND PDBSELECT	36
4.2 LUA OF ENZYMES, NON-ENZYMES, DISPROT, PDBSELECT	38
4.3 LUA OF ENZYMES, NON-ENZYMES	39

LIST OF FIGURES

3.1 PHYLOGENETIC TREE	37
4.1 COMPARISON OF LUA OF 65 COGS AND PDBSELECT	37
4.2 COMPARISON OF LUA OF E, NE, DISPROT OVER PDBSELECT	39
4.3 COMPARISON OF LUA OF E WITH 65 COGS	40
4.4 COMPARISON OF LUA OF E WITH NE.....	41
4.5 COMPARISON OF LUA OF E,NE (DM).....	41
4.6 COMPARISON OF LUA E(DM) WITH NE(DM).....	42
4.7 COMPARISON OF LUA OF 65 COGS(DM) WITH NE(DM)	43
4.8 AVERAGE FREQUENCY FOR ‘CFWY’(PDB).....	44
4.9 AVERAGE FREQUENCY FOR ‘CFWY (MONOMERS).....	45
4.10 AVERAGE FREQUENCY FOR ‘CFWY’ (SWISSPROT).....	46
4.11 PROBABILITY OF NOT HAVING A RESIDUE (PDB).....	46
4.12 PROBABILITY OF NOT HAVING A RESIDUE (MONOMERS)	47
4.13 PROBABILITY OF NOT HAVING A RESIDUE (SWISSPROT)	48
4.14 AGE V/S PROBABILITY (PDB)	49
4.15 AGE V/S PROBABILITY (MONOMERS)	50
4.16 AGE V/S PROBABILITY (SWISSPROT).....	50
4.17 BOOTSTRAPPING	51
4.18 COMPARISON OF PROBABILITY IN E,NE,65 COGS	52
4.19 COMPARISON OF PROBABILITY IN E,NE,65 COGS,MONOMER ...	53

ACKNOWLEDGEMENTS

The graduate education at School of Informatics at Indiana University – Purdue University – Indianapolis has been an excellent experience. Now I would like to take the opportunity to express my gratitude to all those who made my graduate studies a memorable experience and made this thesis possible.

I would like to thank my advisor Dr. Pedro Romero for his continuous guidance and support during the course of my thesis. His constant feedback and encouragement helped me lay a strong foundation for the thesis. I would also thank him for his encouragement in learning and exploring new concepts.

I would also like to thank Dr. Vladimir N Uversky and Dr. Keith A Dunker for being a part of my advisory committee and providing valuable inputs towards the thesis. Their meticulous efforts to review my work are greatly appreciated. Also, I would like to add my thanks to the faculty and staff of the Department of Bioinformatics for their cooperation.

I would like to thank my family for their unstinting support and for providing an opportunity to study at IUPUI. Last but not the least; I would like to thank all my teammates of the project, Christopher J Oldfield and Wai Chan for their cooperation and all my friends, Pradeep, Premchand, Vamshi, Amar, Srikant and Rahul for their encouragement and assistance throughout the course of my study.

ABSTRACT

SAI HARISH BABU KARNE

INTRINSIC DISORDER AND PROTEIN EVOLUTION: AMINO ACID COMPOSITION OF PROTEINS IN LAST UNIVERSAL ANCESTOR

All twenty amino acids did not appear simultaneously in nature. Instead some of them appeared early, while others were added into the genetic code later. The amino acids that were formed by Miller (1953) are suggested to have appeared early in evolutionary history, and the amino acids associated with codon capture developed late in the course of evolution. The chronological order of appearance of the amino acids proposed by Trifonov (2000) was G/A, V/D, P, S, E/L, T, R, N, K, Q, I, C, H, F, M, Y, W. According to Romero et al. (1997) amino acids G, D, E, P and S are disorder-promoting residues and C, F, W and Y are order-promoting residues this means that the early or the ancient amino acids were disorder promoting and the order promoting residues came late into the genetic code. These observations led to the hypothesis that the first proteins, which were comprised of the early amino acids only, were disordered, and, furthermore, that the appearance of the late amino acids and the appearance of the structural proteins were concurrent. Software developed by Brooks et al. (2004) to find the amino acid composition of the LUA (Last Universal Ancestor) was used to test this hypothesis. For this work, the Clusters of Orthologous Groups of proteins (65 COGs) were split into enzymes and non-enzymes. It was found that intrinsic disorder was abundant in both the groups of proteins, with non enzymes being much more disorder than enzymes. Further analysis was done to check for the frequency of the modern amino acids C, F, W, and Y in the Protein data bank (PDB) and Swissprot.